DATABASE
The Journal of Biological Databases and Curation

Database, 2020, 1–10
doi: 10.1093/database/baaa007
Original article

Original article

uORFlight: a vehicle toward uORF-mediated translational regulation mechanisms in eukaryotes

Ruixia Niu[1,†], Yulu Zhou[1,†], Yu Zhang[1], Rui Mou[1], Zhijuan Tang[1], Zhao Wang[1], Guilong Zhou[1], Sibin Guo[2], Meng Yuan[3] and Guoyong Xu[1,*]

[1]State Key Laboratory of Hybrid Rice, Institute for Advanced Studies (IAS), Wuhan University, Wuhan, Hubei 430072, China, [2]Guangxi Key Laboratory of Rice Genetics and Breeding, Rice Research Institute, Guangxi Academy of Agricultural Science, Nanning, Guangxi 530007, China and [3]National Key Laboratory of Crop Genetic Improvement, National Centre of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan, Hubei 430070, China

*Corresponding author: Tel: +8627-68758136; Fax: +8627-68758136; E-mail: guoyong.xu@whu.edu.cn
†These authors contributed equally to this work.

Citation details: Niu,R., Zhou,Y., Zhang,Y. *et al.* uORFlight: a vehicle toward uORF-mediated translational regulation mechanisms in eukaryotes. *Database* (2020) Vol. 2020: article ID baaa007; doi:10.1093/database/baaa007

Received 13 June 2019; Revised 9 January 2020; Accepted 16 January 2020

Abstract

Upstream open reading frames (uORFs) are prevalent in eukaryotic mRNAs. They act as a translational control element for precisely tuning the expression of the downstream major open reading frame (mORF). uORF variation has been clearly associated with several human diseases. In contrast, natural uORF variants in plants have not ever been identified or linked with any phenotypic changes. The paucity of such evidence encouraged us to generate this database-uORFlight (http://uorflight.whu.edu.cn). It facilitates the exploration of uORF variation among different splicing models of Arabidopsis and rice genes. Most importantly, users can evaluate uORF frequency among different accessions at the population scale and find out the causal single nucleotide polymorphism (SNP) or insertion/deletion (INDEL), which can be associated with phenotypic variation through database mining or simple experiments. Such information will help to make hypothesis of uORF function in plant development or adaption to changing environments on the basis of the cognate mORF function. This database also curates plant uORF relevant literature into distinct groups. To be broadly interesting, our database expands uORF annotation into more species of fungus (*Botrytis cinerea* and *Saccharomyces cerevisiae*), plant (*Brassica napus*, *Glycine max*, *Gossypium raimondii*, *Medicago truncatula*, *Solanum lycopersicum*, *Solanum tuberosum*, *Triticum aestivum* and *Zea mays*), metazoan (*Caenorhabditis elegans* and *Drosophila melanogaster*) and vertebrate (*Homo sapiens*, *Mus musculus* and *Danio rerio*). Therefore, uORFlight will light up the runway

toward how uORF genetic variation determines phenotypic diversity and advance our understanding of translational control mechanisms in eukaryotes.

Database URL: http://uorflight.whu.edu.cn/

Introduction

Gene expression must be tightly regulated at transcription, translation and post-translation levels. The imperfect correlation between protein abundance and mRNA levels suggests translation efficiency regulated by translational control as one of the determinants of protein outputs from variable mRNA inputs. This layer of regulation is mediated by the cooperative action between different mRNA elements and *trans*-acting factors (1). Upstream open reading frames (uORFs) are among the mRNA elements that can confer precise control of protein translation.

A uORF initiation codon resides upstream of the coherent mORF, and will be first encountered by 43S scanning ribosome (including 40S ribosomal subunit and eIF2 ternary complex). Sequentially, 60S subunit joins in and reconstitutes 80S ribosome for uORF translation elongation, after which the 40S and 60S are disjointed and 40S may remain associated with mRNA. Therefore, usually uORF translation is prioritized over mORF, leading to hindered translation of the mORF. Only in situations where the remaining 40S ribosomes regain fresh eIF2 ternary complex and other unknown reinitiation factors, or when uORF initiation codon is bypassed by the scanning ribosome, the downstream mORF has the chance to be translated (Figure 1a). The former situation is termed as reinitiation and the latter as leaky scanning, two mechanisms that have been accepted as an explanation of limited mORF expression under normal growth and developmental conditions (2, 3). Therefore, genome editing of uORF to remove its translational suppression of a key enzyme in vitamin C biosynthesis engineers oxidation stress tolerant and antioxidant metabolite enriched plants (4). Most importantly, a uORF can confer selective mORF translation in response to a wide range of cellular stimuli, such as metabolite and ion homeostasis, hormone changes, environmental signals and immune induction (See uORF references on our website). This tight and temporal regulation pattern fine-tunes the translation efficiency of mORFs and thus guarantees appropriate protein quantity and quality for adaption to different physiological conditions. Because of those unique features, we have successfully utilized uORF-mediated translational control in engineering disease resistant plants without fitness costs by restricting toxic resistance protein translation under normal conditions but allowing transient induction under pathogen infection conditions (5).

However, once uORF-mediated precision control has been challenged by genetic variation or mis-regulation, it causes human diseases (6, 7). By 2009, 509 human genes had been identified with polymorphic uORFs and some of them have been experimentally associated with different human diseases, including malignancies, metabolic or neurologic disorders, and inherited syndromes. This trend became more striking recently with more genomic variation data released and analyzed (8, 9). In contrast, natural variation of plant uORFs has not yet been investigated, even though there are abundant publicly accessible genetic and phenotypic variation data, especially for model organisms Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa* L.). Since the release of the Arabidopsis reference genome of accession Col-0 in 2000, the rapid development of sequencing technology has bolstered genome-wide association studies (GWAS) by linkage disequilibrium of interesting phenotypic traits with the most probable genetic variation, particularly using the genome sequences of 1135 accessions from the 1001 Genomes Project (10–16). Genetic variation of rice has long been used for molecular breeding and recent re-sequencing of a large set of rice accessions, especially from the 3000 Rice Genomes Project, generated a wealth of genetic variation for the discovery of useful alleles for agronomic trait improvement (17–25).

It is noteworthy that previous genotype–phenotype association studies mainly focused on protein coding regions, while it is becoming more evident that the *cis*-element variation weighs a lot in determining phenotypic variation, such as variation of the promoter regions and alternative transcription starting sites in changing fruit sizes and light responses, respectively (26, 27). However, there is still less attention that has been paid on the variation of mRNA regulatory elements, such as uORFs. In this study, we used public resources to identify uORF variation for further experimental verification of phenotypic diversity mediated by translational control.

Methods

A uORF is defined as the presence of an initiation codon in an annotated mRNA 5′ leader region and can be categorized into ‘Types 1–3’ based on the positions of uORF stop codon, with Type1 in 5′ leader, Type2 in mORF coding region and Type3 shared with mORF (also known as an mORF N-extension). It is obvious that reinitiation

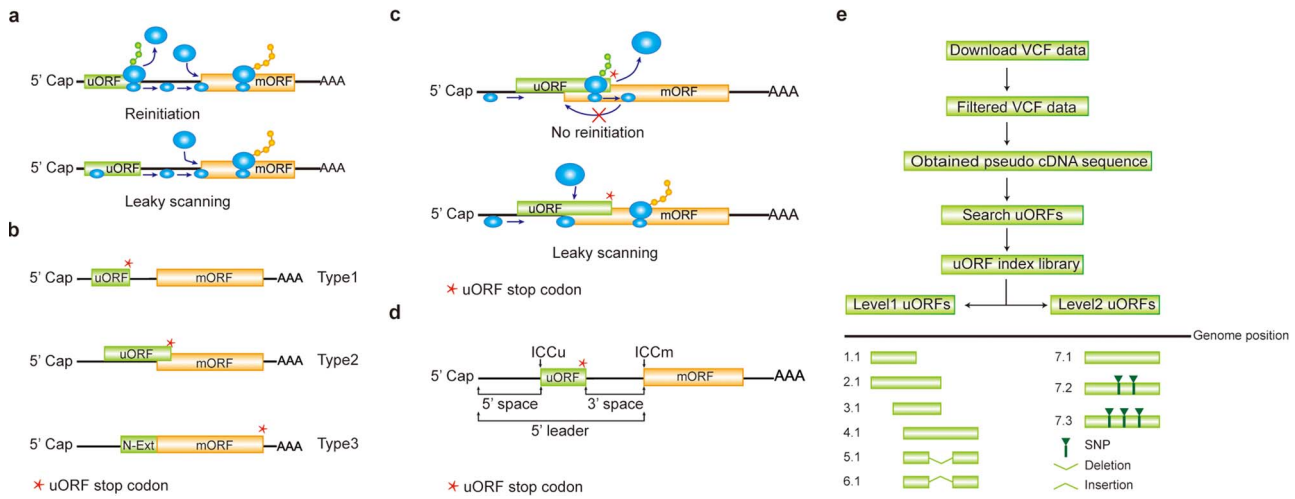

Figure 1. Computation of uORF variation. (**a**) Reinitiation and leaky scanning models. In the reinitiation model, the 80S ribosomal subunit will separate after translating the uORF and the 40S subunit remains associated with mRNA, regaining fresh eIF2 ternary complex and other unknown reinitiation factors to translate the mORF. In the leaky scanning model, the uORF initiation codon is bypassed by the scanning complex, which will ignore the uORF and translate the mORF. (**b**) uORF types. uORFs are divided into Types1–3 with respect to the position of uORF stop codon relative to the mORF. N-Ext, N extension. (**c**) Type2 uORF-controlled mORF translation is only favored by leaky scanning. Overlap between the Type2 uORF and mORF makes reinitiation of the mORF impossible after translation of uORF. (**d**) uORF positional information on cDNA. The mORF is flanked by the 5′ leader and 3′ UTR (3′ untranslated region). 5′ and 3′ space are used to describe the distance from Cap to uORF AUG and from uORF stop codon to mORF AUG, respectively. The sequence from −3 to +4 relative to the AUG initiation codon (A as +1) corresponding to the Kozak consensus (A/GCCAUGG) position is termed as initiation codon context (ICC) with ICCu and ICCm for uORF and mORF, respectively. (**e**) Workflow to identify uORF variants. VCF files are downloaded from public databases and filtered as described in the **Method**. Pseudo cDNA sequences are generated by replacing all the splicing models with filtered variants. After uORFs identification, their relative positions on the pseudo cDNAs are transferred into absolute positions in the reference genome to create an index library. uORFs are pairwise compared and those uORFs with the same genomic positions and sequences are assigned the same index number to indicate no uORF changes. Otherwise, two leveled uORF identifiers (Level1.Level2) are used to describe the variation. Level1 indicates major differences that cause uORF creation, loss or length changes due to SNPs or INDELs (insertion and deletion). Level2 indicates minor differences that lead to nucleotide and/or amino acid substitution due to SNPs. Continuous index numbers starting with number 1.1 are assigned in the prioritized orders of uORF ATG occurrence in the reference genome, uORF length (shorter), deletion and insertion at Level1, and of SNP number (fewer) at Level2.

is impossible for translation of Type2 uORF-controlled mORFs (Figure 1b and c). Hereafter, ORF of both uORF and mORF means that AUG is used as the initiation codon, unless specifically stated. We use the term 5′ leader sequence instead of 5′ UTR, considering the peptide-coding potential of uORFs (28, 29). The sequence from −3 to +4 relative to AUG initiation codon (A as +1) corresponding to Kozak consensus (A/GCCAUGG) position is termed as initiation codon context (ICC) with ICCu and ICCm for uORF and mORF, respectively (Figure 1d).

We chose the Arabidopsis Col-0 accession (Ensemble V39; Araport11) and rice Nipponbare cultivar (MSU V7) as reference genomes for dicot and monocot uORF analysis, respectively. Arabidopsis representative gene models (27 445 in total; Supplementary Table 1 in **Download** menu) of the nuclear coding proteins contain 26 713 genes using TAIR10 representative gene annotation file and 732 genes using their '0.1' splicing models. Nipponbare representative gene models (38 860 in total; Non-TE Loci) are defined using their smallest numbered models (38 618, 221, 16 and 5 genes using the '0.1', '0.2', '0.3' and '0.4' splicing models, respectively; Supplementary Table 2 in **Download** menu). To calculate uORF variation, we downloaded VCF (variant call format) files of Arabidopsis 1135 accessions from the 1001 Genomes Project and rice 3k varieties from the 3000 Rice Genomes Project (10, 21, 25, 30). We filtered single nucleotide polymorphisms (SNPs) and insertion/deletions (INDELs) with low quality indicated in the VCF files and used the alleles with frequency over 90% as suggested (30). We further removed genes with variants affecting the annotated initiation codon of their mORFs because we need mORF initiation codon as a fixed coordinate for locating uORFs.

We replaced all the splicing models with filtered variants and searched uORFs in all the accessions. We then transferred the relative positions of uORFs on the different transcripts into absolute positions in the reference genome and assigned continuous index numbers starting with number 1.1 in the order of uORF ATG occurrence in the genome. Those uORFs with the same genomic positions and sequences are assigned the same index number to indicate no uORF changes. Otherwise, two leveled uORF identifiers (Level1.Level2) are used to describe the variation. Level1 indicates major differences that cause uORF

creation, loss or length changes due to SNPs or INDELs. Level2 indicates minor differences that lead to nucleotide and/or amino acid substitution due to SNPs (Figure 1e). We grouped Arabidopsis accessions on the basis of the latitude at which they were collected (15-degree interval; Supplementary Table 1 in **Download** menu). The frequency of an individual uORF in Arabidopsis was calculated based on its occurrence in the total population and in different latitude ranges. The frequency in rice was calculated as its occurrence in total population and nine subspecies (21). The associated SNP or INDEL identifiers are also recorded along with uORF variants and are searched against Arabidopsis GWAS database (http://1001genomes.org/), and rice GWAS (http://ricevarmap.ncpgr.cn/v2/) and quantitative trait locus (QTL) databases (https://archive.gramene.org/db/qtl/qtl_display?species=Oryza%20sativa).

MySQL database schema was used for uORF information storage and a user-friendly PHP web interface was designed to query and download. Gene Ontology (GO) analysis was done using Omicshare online tools (http://www.omicshare.com/) with the default setting. The uORF annotation of the other species can be found on the website.

Results and discussion

With the recent recognition of the significance of uORFs within distinct physiological contexts, the following functionalities will help the community quickly overview the progress in this area and find out uORF variation to link with phenotypic diversity at the individual gene level (comparing uORF variation among different splicing models) and at the population level (comparing uORF variation among different accessions).

uORF in the reference genomes

uORFs are becoming increasingly attractive because of their capacity to fine-tune translation and respond accurately to distinct extracellular and intracellular stimuli. However, the current understanding of uORFs is based on a small number of case studies. In an attempt to provide guidelines, we investigated natural patterns in uORF types, length distribution and ICC of Arabidopsis and rice representative gene models. uORF-containing genes are more prevalent in Arabidopsis (48.45%), and the lower frequency in rice genes (20.65%) may arise from current incomplete 5′ leader annotation. Their prevalence is mostly due to overrepresentation of Type1 uORFs, which account for 90.79% and 87.63%, in contrast to only 9.16% and 12.17% for Type2 uORFs of Arabidopsis and rice, respectively. The Type3 uORFs are the least common (19 uORFs in 17 Arabidopsis genes, and 74 uORFs in 41 rice genes), and they may give rise to N-extension and are likely to alter protein activities or molecular localizations as reported (31). Type2-containing genes tend to occur along with Type1 uORFs, and the significance of their co-existence needs further investigation (Figure 2a).

uORF-containing genes are categorized into uORF-free (no uORF), Type1-only and Type2 groups (at least containing one Type2 uORF). GO analysis shows clearly different enriched terms among those groups in both Arabidopsis and rice, suggesting that uORF type attributes could have an impact on different functional groups of genes (Figure 2b; Supplementary Tables 1 and 2 in **Download** menu). This assumption appears to be reasonable because both leaky scanning and reinitiation mechanisms could overcome Type1 uORF inhibition, while only leaky scanning can overcome Type2 uORF inhibition (Figure 1a and b), and leaky scanning seems to be of low efficiency in both animals and plants (32, 33). Importantly, uORF-containing genes are enriched in more specific groups than uORF-free genes. In the uORF-containing genes, GO terms including key words such as ‘response’, ‘regulation’ and ‘signaling’ are more frequently observed. Moreover, even though rice has fewer enriched GO-terms, most of them (89.47%) are shared with GO-terms enriched for Arabidopsis uORF-containing genes. The 11 shared GO-terms for Type1-only and Type2 include ‘signaling’ and ‘cellular response to stimulus’ (Figure 2c). These findings are consistent with our current knowledge of a few functional uORFs that are involved in ‘signaling’ and ‘cellular response to stimulus’, such as starvation responsive gene GCN4 in yeast (34), hypoxia and endoplasmic reticulum stress-responsive gene ATF4 in mammalians (35) and sucrose responsive gene bZIP11 and immune responsive gene TBF1 in Arabidopsis (36, 37). All these results suggest that uORFs are more frequently used by specific adaption-related groups and these groups are shared between Arabidopsis and rice.

Next, we examined uORF length distribution and found that Type2 uORFs are on average longer than Type1, and that rice uORFs are generally longer than those of Arabidopsis (Figure 2d). However, uORF length appears not to be a definitive parameter for prediction of functional uORF on translational control. A minimal Type1 uORF consisting of an AUG and a stop codon has been shown to be sufficient for translational inhibition of three boron (B)-related genes and also sufficient for translational responsiveness to low B stimulation (38). Such short uORFs (6.79%) are more commonly found among Arabidopsis Type1 uORFs and must require other synergistic *cis*-elements to allow specifically translational responsiveness to a stimulus. We then asked whether uORF surrounding sequences display some informative patterns. We found that only 0.45% ICCu and

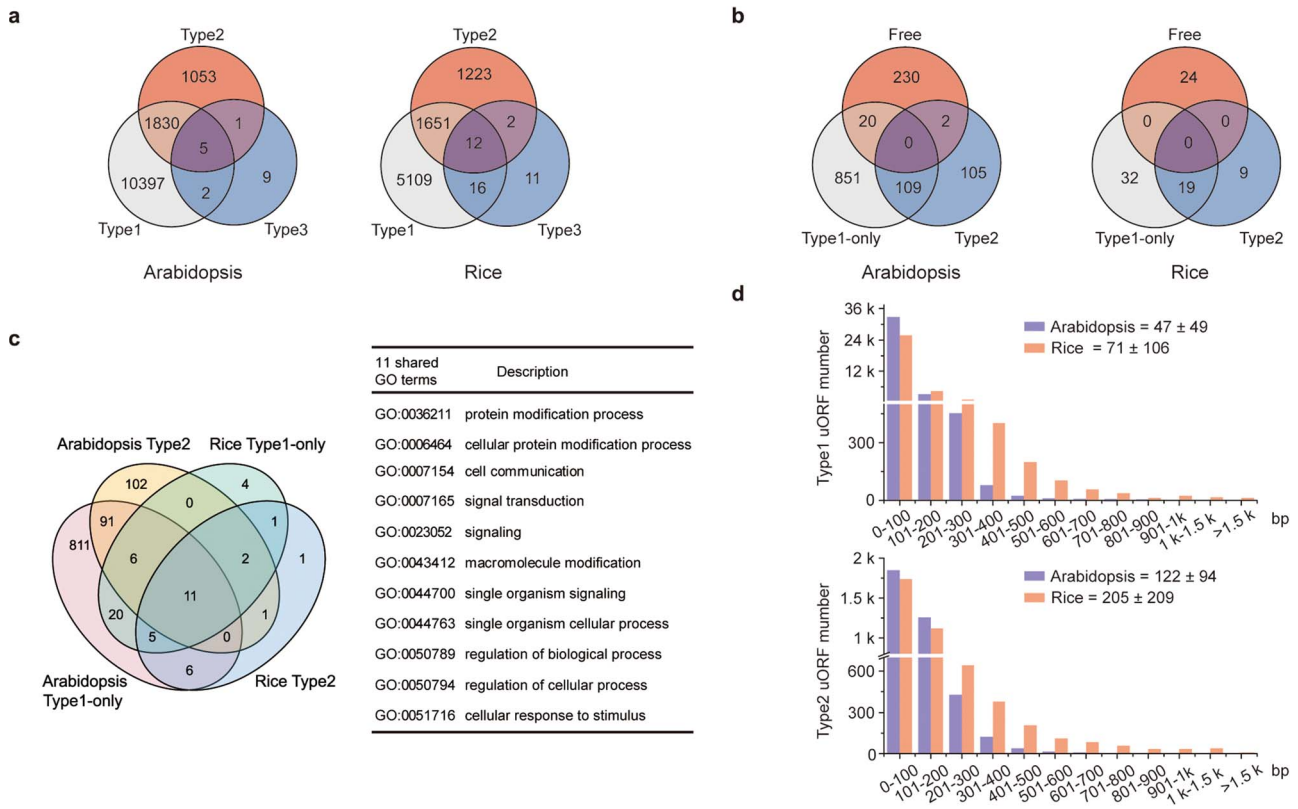

Figure 2. Characterization of uORFs in Arabidopsis and rice. (**a**) Venn diagram of uORF-containing genes with different uORF types. (**b**) Venn diagram of enriched GO terms of uORF-free-, Type1-only- and Type2- uORF-containing genes in Arabidopsis and rice. GO terms of $P < 0.05$ are used. (**c**) Venn diagram to show GO terms shared in Arabidopsis and rice uORF-containing genes. 51 out of 57 enriched GO terms of rice uORF-containing genes are also found in Arabidopsis. Eleven GO terms enriched in both Type1-only and Type2 uORF-containing genes of Arabidopsis and rice are detailed in the table. (**d**) Length distribution of Type1 and Type2 uORFs. Insert number shows Mean ± SD of uORF length (bp). The representative gene models of Arabidopsis reference accession Col-0 and rice reference cultivar Nipponbare are used for analysis.

2.99% ICCm in Arabidopsis and 1.93% ICCu and 11.88% ICCm in rice contain Kozak consensus sequences, suggesting that more variable contexts are used to flexibly tune ORF translation in nature of plants. Enrichment analysis of ICCu and ICCm did not uncover any obvious enriched sequence, suggesting the feasibility of engineering tailored protein expression using ICC variants in plants.

uORF variation at the population level

uORF variation has been studied at the genome-wide scale using human SNPs and the results provide clear evidence of its association with genetic diseases. However, these studies only assessed effects of a single SNP that result in uORF initiation codon creation or stop codon loss (8, 9). We analyzed uORF variation by considering the integral effect of all homozygous SNPs and INDELs in each plant accession and found that 54.17% of Arabidopsis uORF variants (64.52% in rice) of the representative gene models are associated with at least two SNPs and/or INDELs. As emphasized above, uORF creation, loss, length changes and type switches are considered as major variation and are defined as Level1 (Figure 1e). Level2 variation is likely to affect uORF function, provided that its nucleotide sequence or encoded peptide is able to cause ribosome stalling or mRNA decay (2, 6, 7). To simply show the significance of uORF variation, we focused on three Level1 variants, which alter uORF type attributes of genes between Type1-only and Type2, between uORF-free and Type1-only, and between uORF-free and Type2 (Figure 3a–c).

We first found that 31.03% and 20.40% of uORF-containing genes have altered uORF type attributes among different splicing models in Col-0 and Nipponbare reference genomes, respectively, suggesting that uORF variation produced by alternative splicing adds translational control to specific processes, such as developmental pathways (Arabidopsis flowering: FCA; stem cell: CLV3), hormonal pathways (jasmonic acid: Arabidopsis JAZ4 and rice JAZ1; auxin: Arabidopsis IAA7/11 and rice IAA4; ethylene: Arabidopsis EIL3; brassinosteroid: Arabidopsis BES1 and rice DWF1), signaling transduction pathways (Arabidopsis CIPK1/17; rice MAPK4/83) and immune responses (Arabidopsis SNC1, RPS5, PBL1 and MLO2; rice Rac1 and Pi21) (Figure 3c and d). It is likely that different splicing

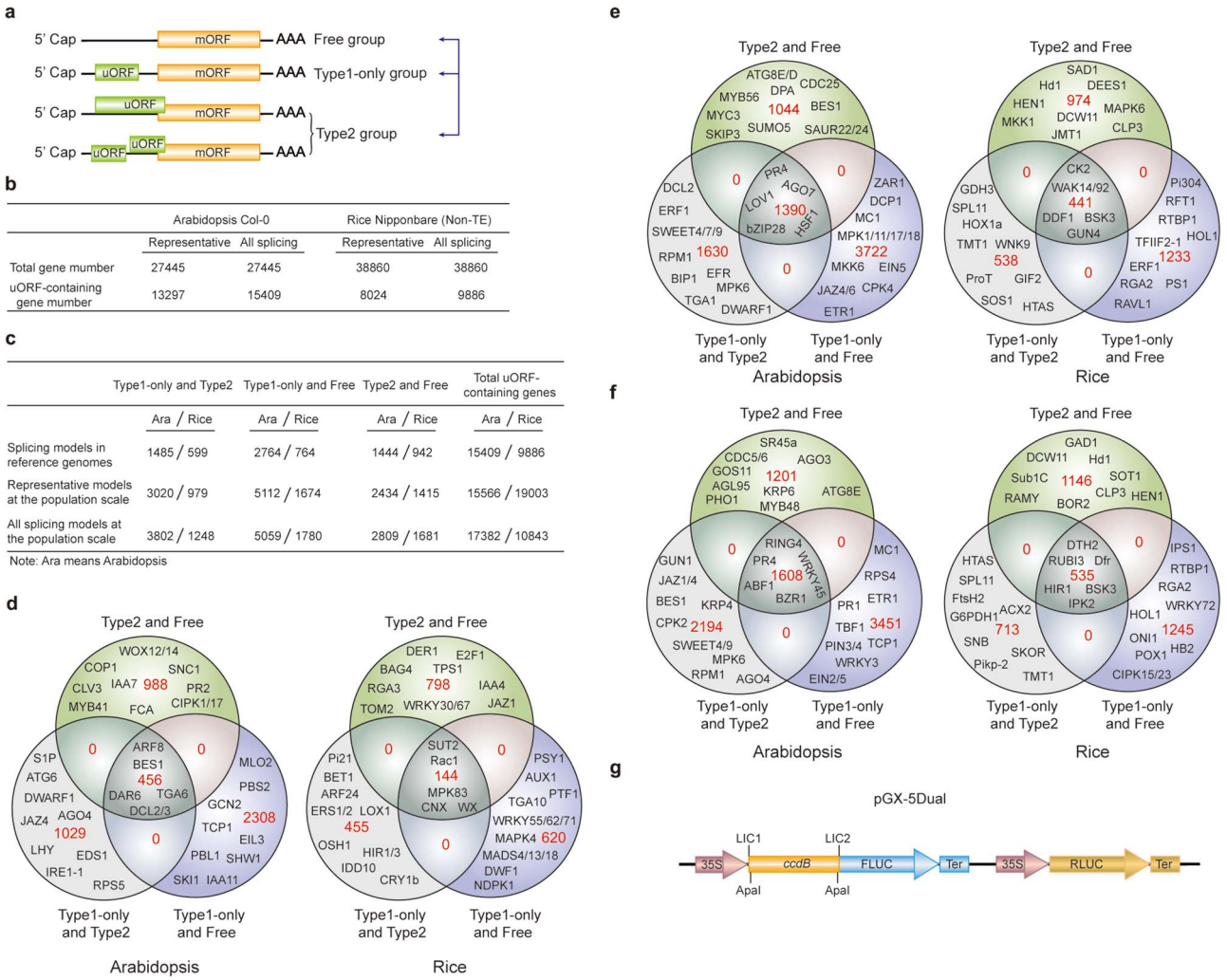

Figure 3. uORF variation. (**a**) Schematic of altered uORF types analyzed in this study. Three Level1 uORF variants are focused by studying changes between Type1-only and Type2, between Type1-only and uORF-free and between Type2 and uORF-free. Genes with both Type2 and Type1 on the same splicing model are grouped as Type2 in this analysis. (**b**) Number of uORF-containing genes in Arabidopsis Col-0 and rice Nipponbare. The representative and all splicing gene models are calculated. Non-TE, non-transposon element. (**c**) Number of three Level1 uORF variants. Analysis is performed among different splicing models of reference genomes (Arabidopsis Col-0 and rice Nipponbare), and among accessions at the population scale using the representative and all splicing gene models. (**d**) Venn diagram of three Level1 uORF variants within different splicing models of Arabidopsis Col-0 and rice Nipponbare. (**e**, **f**) Venn diagram of three Level1 uORF variants of the representative gene models (e) and all splicing gene models (f) at the population level. Example gene names are shown in different regions. (**g**) Schematic of dual-luciferase vector pGX-5Dual. 5′ leader containing different uORF variants are cloned via LIC method to control the translation of firefly luciferase (FLUC). ApaI is used to linearize the vector. Comparison of the ratio of FLUC activity and mRNA level to the internal control RLUC expressed from the same vector will indicate the effects of uORF variants on mRNA stability and translation efficiency.

models give rise to both uORF and mORF variants, of which uORF variants provide distinct translational control capacity and mORF variants provide variable protein activity. Their concurrent and synergistic actions may allow the most optimized adaption of mORF-encoded protein activity to changing developmental and environmental cues. This finding also emphasizes the significance of taking into accounts of uORF variation during mRNA splicing studies, which will be of great promise and interest in the future.

We next identified the three Level1 variants of different accessions at the population level by analyzing the representative gene models (Figure 3e). Alternatively, we also conducted the analysis by considering all splicing models of a gene, and any splicing model of one accession different from the other accessions are counted having Level1 uORF variants (Figure 3f). Significantly, these two different ways exemplify the importance of uORF variation on key genes with different cellular functions, such as hormonal genes (Arabidopsis JAZ1/4/6, EIN2/5,ETR1 and BES1) and immune genes (Arabidopsis EFR and MPK6). Most importantly, many rice agronomic traits-related genes are found to have uORF variants among different accessions, sug-

gesting that uORF is utilized to translational control of specific agronomic traits, such as Hd1 for heading date and yield, Sub1C for submergence tolerance and Pikp-2 for *Magnaporthe oryzae* resistance (Figure 3e and f).

In an attempt to associate uORF variants with known phenotypes, we suggest two methods. First, we reason that if the causal SNPs and INDELs of uORF variants are located within known QTL regions or the causal SNPs are consistent with GWAS SNPs, it is possible that uORF variants contribute to the phenotype linked to the corresponding QTL or GWAS. It should be noted that one QTL region has more than uORF-associated SNPs and INDELs, and thus uORF variants within this region only suggest the possibility of translational control of this QTL phenotype. Further, the current GWAS databases use SNPs as genetic variants to perform association studies. Methods using INDELs are still under active development (39). On the other hand, many variables, such as population size, may cause GWAS false positive or biased associations (40). Last, GWAS data are limited especially for rice and we did not detect their association with uORF variants in rice. Nevertheless, considering uORF variants are responsible for 30–80% protein abundance changes in human (41), we highly recommend directly performing a simple dual-luciferase assay to compare the translation efficiency of uORF variants by Agrobacterium-mediated transient expression in *Nicotiana benthamiana* (42). Accordingly, we provide a dual-luciferase vector (pGX-5Dual) along with this database. Users can easily clone 5′ leaders containing different uORF alleles to the 5′ of firefly luciferase (FLUC) through LIC (ligation-independent cloning) technology (43). Comparison of the ratio of FLUC activity and mRNA level to the internal control Renilla luciferase (RLUC) expressed from the same vector will indicate the effects of uORF variants on mRNA stability and translation efficiency. The difference will provide hypothesis for the association of uORF variation with its physiological roles deduced from its cognate mORF function (Figure 3g).

Conserved peptide uORF information

With recent development of ribosome footprint and proteomics, more uORFs have been found to encode polypeptides. Current annotation rule defines mORF with the longest ORF whose encoded protein has homologs in other species. Analysis of evolutionary conservation suggests the existence of conserved peptide uORFs (CPuORFs) (44–50) and 97 non-redundant AUG-initiated CPuORFs in Arabidopsis have been confirmed here. Therefore, a better annotation of mORF needs to consider ORF length, peptide conservation, ribosome binding characteristics and peptide abundance. It is suggested that most CPuORFs confer peptide sequence-dependent regulation in a *cis* manner, as metabolite receptors or sensors that function in the ribosome exit tunnel by stalling the ribosome and preventing reinitiation (2). There are also exceptional reports on uORFs functioning in *trans*-regulation (51, 52). In '**uORF view**' menu, we collected all Arabidopsis genes containing CPuORF and also provided their rice homologs (Supplementary Table 3 in **Download** menu). This information will advance our understanding of uORF sequence-dependent regulation by facilitating the study of CPuORFs.

Plant uORF reference curation

Eukaryotic uORF-related literatures to year 2013 have been curated in uORFdb through the Boolean search for key words in the NCBI PubMed database (53). To help the community get the latest view of the progress in plant uORF research, we manually curated all the relevant references on the basis of our knowledge and categorized the references into Case study, Mechanism study, Practical study, Genome-wide study and Review. Users are invited to help us complete this section if missing or inappropriate references are found. Clicking the reference link will direct the users to the associated PubMed page or journal page.

Customized uORF analysis

In the current database, we only processed uORFs with AUG as the initiation codon. With the rapid development of ribosome footprinting, mounting evidence suggests the usage of non-AUG initiation codons in translational control (54). Those non-AUG-initiated ORF-encoded peptides are also detectable by mass spectrometry (28). However, non-AUG-initiated uORFs may function in an opposite way to AUG uORFs in translational control (54). This nuance remains poorly understood and systematic variation analysis of non-AUG-initiated uORFs will be included when more information is available in plants. Therefore, users who are interested in non-AUG studies or uORF studies in species not included in our database are encouraged to use our searching tool under the menu of **Tools** to obtain the basic uORF information by inputting different uORF initiation codons, cDNA and CDS sequences separately.

Navigation of this website

The structure and the main functionalities of our website are depicted in Figure 4. From **uORF view** menu, users can browse and search uORF information for all splicing models in the reference genomes, including dicot **Arabidopsis** Col-0, monocot **Rice** Nipponbare and **other species** including *Botrytis cinerea* (gray mold),

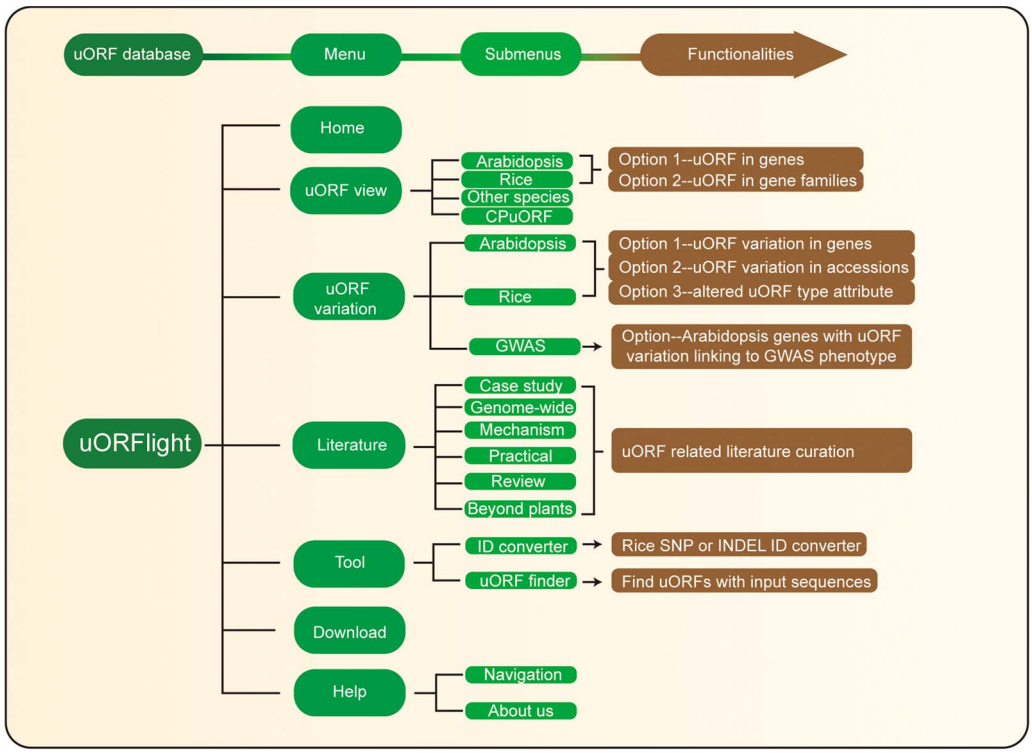

Figure 4. Database structure. **Home** menu contains the background information of uORFlight database including organisms, and definition of uORF attributes and variants. **uORF view** menu has four submenus to browse and search uORF, including in the reference genomes of Arabidopsis Col-0, rice Nipponbare and other species. CPuORF is also included in this menu. In Arabidopsis and rice submenus, Option 1 and Option 2 are provided to individually and bulk retrieve uORF information, respectively. **uORF variation** menu is used to compare uORF variation among different splicing models in the reference genome (Option 1) or among the selected accessions (Option 2) and to bulk retrieve genes with altered uORF types (Option 3). **Literature** menu curates plant uORF relevant literature into distinct groups. **Tool** menu provides ID converter and uORF finder with the former to transform SNP and INDEL variation identity used in different external databases, and with the later to search ATG or non-ATG initiated uORFs in a given cDNA sequence. **Help** menu contains **Navigation** submenu to explain the main conclusion on each result page.

Saccharomyces cerevisiae (budding yeast), *Brassica napus* (rapeseed), *Glycine max* (soybean), *Gossypium raimondii* (cotton), *Medicago truncatula* (grass), *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Triticum aestivum* (wheat), *Zea mays* (corn), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Homo sapiens* (human), *Mus musculus* (mouse) and *Danio rerio* (zebrafish). In **Arabidopsis** and **Rice** submenus, **Option 1** and **Option 2** are provided with the former to return uORF information for a gene, and with the later to bulk retrieve uORF information for a popular gene family, such as kinases, transcription factors (TFs) and nucleotide-binding leucine rich repeat proteins (NLRs; immune genes).

From Arabidopsis and rice submenus of **uORF variation** menu, users can compare uORF variation among different splicing models in the reference genomes using **Option 1** or among the selected accessions using **Option 2**. In the returning page, uORF distribution in different splicing models and in different groups of accessions (Arabidopsis: latitude ranges; rice: subspecies) is plotted. Detailed information including the associated QTL can be accessed in the downloaded tables. In addition, we also provide **Option 3** for users to bulk retrieve genes with altered uORF types between Type1-only and Type2, between uORF-free and Type1-only, and between uORF-free and Type2. GWAS submenu will return Arabidopsis genes with uORF variants associated with selected GWAS phenotypes. However, the limited GWAS information of rice does not find any genes with uORF variants linked to known GWAS phenotypes. To help users interpret the searching results, we also include descriptive paragraphs to explain the main conclusion on each result page in our Navigation submenu under the Help menu.

Future direction

uORFs are common in eukaryotes and information from more organisms will be useful additions to our database in the future. uORFs may encode functional peptides to act in either *trans* or *cis* manners, and this information will need to be evaluated by the combination of ribosome footprinting and mass spectrometry data, which will be integrated as it becomes available. In addition, uORFs are RNA *cis*-elements that require *trans*-acting factors to reg-

ulate translation. Meanwhile, co-regulatory *cis*-elements, such as the R-motif identified in our previous study, may account for uORF regulation specificity and diversity. Information about regulatory *trans*-acting factors and co-acting *cis*-element variation will be incorporated into the database progressively. uORF variation due to alternative transcription starting sites is exemplified in light responses, and systemic computation of the effect of alternative transcription starting sites on uORFs variation will be also considered upon more experimental data available. Furthermore, a uORF calculator will be developed and installed to predict the regulatory power of natural or synthetic uORFs for tailored protein expression after machine learning of large experimental data is achieved.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank Sophia Zebell and Paul J. Zwack at Duke University for the comments.

Funding

National Natural Science Foundation of China (31822042) to M.Y.; start-up fund from Wuhan University to G. Xu.

Conflict of interest. None declared.

References

1. Merchante,C., Stepanova,A.N. and Alonso,J.M. (2017) Translation regulation in plants: an interesting past, an exciting present and a promising future. *Plant J.*, **90**, 628–653.
2. von Arnim,A.G., Jia,Q. and Vaughn,J.N. (2014) Regulation of plant translation by upstream open reading frames. *Plant Sci.*, **214**, 1–12.
3. Jackson,R.J., Hellen,C.U. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.
4. Zhang,H., Si,X., Ji,X. *et al.* (2018) Genome editing of upstream open reading frames enables translational control in plants. *Nat. Biotechnol.*, **36**, 894.
5. Xu,G., Yuan,M., Ai,C. *et al.* (2017) uORF-mediated translation allows engineered plant disease resistance without fitness costs. *Nature*, **545**, 491–494.
6. Somers,J., Poyry,T. and Willis,A.E. (2013) A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.*, **45**, 1690–1700.
7. Barbosa,C., Peixeiro,I. and Romao,L. (2013) Gene expression regulation by upstream open reading frames and human disease. *PLOS Genet.*, **9**, e1003529.
8. Calvo,S.E., Pagliarini,D.J. and Mootha,V.K. (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. USA*, **106**, 7507–7512.
9. Whiffin,N., Karczewski,K.J., Zhang,X. *et al.* (2019) Characterising the loss-of-function impact of 5′ untranslated region variants in 15,708 individuals. bioRxiv, 543504.
10. The 1001 Genomes Consortium, Alonso-Blanco,C., Andrade,J. *et al.* (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, **166**, 481–491.
11. Ossowski,S., Schneeberger,K., Clark,R.M. *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.*, **18**, 2024–2033.
12. Schneeberger,K., Ossowski,S., Ott,F. *et al.* (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 10249–10254.
13. Cao,J., Schneeberger,K., Ossowski,S. *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.*, **43**, 956–963.
14. Long,Q., Rabanal,F.A., Meng,D. *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.*, **45**, 884–890.
15. Gan,X., Stegle,O., Behr,J. *et al.* (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, **477**, 419–423.
16. Togninalli,M., Seren,U., Meng,D. *et al.* (2018) The AraGWAS Catalog: a curated and standardized Arabidopsis thaliana GWAS catalog. *Nucleic Acids Res.*, **46**, D1150–D1156.
17. Yu,J., Hu,S., Wang,J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
18. Goff,S.A., Ricke,D., Lan,T.H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, **296**, 92–100.
19. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
20. Huang,X., Zhao,Y., Wei,X. *et al.* (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.*, **44**, 32–39.
21. Wang,W., Mauleon,R., Hu,Z. *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.
22. 3,000 rice genomes project (2014) The 3,000 rice genomes project. *Gigascience*, **3**, 7.
23. Wing,R.A., Purugganan,M.D. and Zhang,Q. (2018) The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.*, **19**:505–517.
24. Zhao,H., Yao,W., Ouyang,Y. *et al.* (2015) RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids Res.*, **43**, D1018–D1022.
25. Mansueto,L., Fuentes,R.R., Borja,F.N. *et al.* (2017) Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Res.*, **45**, D1075–D1081.
26. Rodríguez-Leal,D., Lemmon,Z.H., Man,J. *et al.* (2017) Engineering quantitative trait variation for crop improvement by genome editing. *Cell*, **171**, 470–480.
27. Kurihara,Y., Makita,Y., Kawashima,M. *et al.* (2018) Transcripts from downstream alternative transcription start sites evade uORF-mediated inhibition of gene expression in Arabidopsis. *Proc. Natl. Acad. Sci. USA*, **115**:7831–7836.
28. Slavoff,S.A., Mitchell,A.J., Schwaid,A.G. *et al.* (2012) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59.

29. Wu,H.L., Song,G., Walley,J.W. and Hsu,P.Y. (2019) The tomato translational landscape revealed by transcriptome assembly and ribosome profiling. *Plant Physiol.*, **181**, 367–380.
30. Alexandrov,N., Tai,S., Wang,W. *et al.* (2015) SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.*, **43**, D1023–D1027.
31. Xu,F., Huang,Y., Li,L. *et al.* (2015) Two N-terminal acetyltransferases antagonistically regulate the stability of a nod-like receptor in Arabidopsis. *Plant Cell*, **27**, 1547–1562.
32. Schleich,S., Strassburger,K., Janiesch,P.C. *et al.* (2014) DENR-MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth. *Nature*, **512**, 208–212.
33. Wang,L. and Wessler,S.R. (1998) Inefficient reinitiation is responsible for upstream open reading frame-mediated translational repression of the maize R gene. *Plant Cell*, **10**, 1733–1745.
34. Hinnebusch,A.G. (2005) Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.*, **59**, 407–450.
35. Baird,T.D. and Wek,R.C. (2012) Eukaryotic initiation factor 2 phosphorylation and translational control in metabolism. *Adv. Nutr.*, **3**, 307–321.
36. Wiese,A., Elzinga,N., Wobbes,B. and Smeekens,S. (2004) A conserved upstream open reading frame mediates sucrose-induced repression of translation. *Plant Cell*, **16**, 1717–1729.
37. Pajerowska-Mukhtar,K.M., Wang,W., Tada,Y. *et al.* (2012) The HSF-like transcription factor TBF1 is a major molecular switch for plant growth-to-defense transition. *Curr. Biol.*, **22**, 103–112.
38. Tanaka,M., Sotta,N., Yamazumi,Y. *et al.* (2016) The minimum open reading frame, AUG-stop, induces boron-dependent ribosome stalling and mRNA degradation. *Plant Cell*, **28**, 2830–2849.
39. Song,B., Mott,R. and Gan,X. (2018) Recovery of novel association loci in *Arabidopsis thaliana* and *Drosophila melanogaster* through leveraging INDELs association and integrated burden test. *PLOS Genet.*, **14**, e1007699.
40. Visscher,P.M., Wray,N.R., Zhang,Q. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am J Hum. Genet.*, **101**, 5–22.
41. Ferreira,J.P., Overton,K.W. and Wang,C.L. (2013) Tuning gene expression with synthetic upstream open reading frames. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 11284–11289.
42. Xu,G., Greene,G.H., Yoo,H. *et al.* (2017) Global translational reprogramming is a fundamental layer of immune regulation in plants. *Nature*, **545**, 487–490.
43. Xu,G.Y., Sui,N., Tang,Y. *et al.* (2010) One-step, zero-background ligation-independent cloning intron-containing hairpin RNA constructs for RNAi in plants. *New Phytol.*, **187**, 240–250.
44. Hayden,C.A. and Jorgensen,R.A. (2007) Identification of novel conserved peptide uORF homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC Biol.*, **5**, 32.
45. Tran,M.K., Schultz,C.J. and Baumann,U. (2008) Conserved upstream open reading frames in higher plants. *BMC Genomics*, **9**, 361.
46. Takahashi,H., Takahashi,A., Naito,S. and Onouchi,H. (2012) BAIUCAS: a novel BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences and its application to the *Arabidopsis thaliana* genome. *Bioinformatics*, **28**, 2231–2241.
47. Vaughn,J.N., Ellingson,S.R., Mignone,F. and Arnim,A. (2012) Known and novel post-transcriptional regulatory sequences are conserved across plant families. *RNA*, **18**, 368–384.
48. Ebina,I., Takemoto-Tsutsumi,M., Watanabe,S. *et al.* (2015) Identification of novel *Arabidopsis thaliana* upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner. *Nucleic Acids Res.*, **43**, 1562–1576.
49. Hsu,P.Y., Calviello,L., Wu,H.-Y.L. *et al.* (2016) Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc. Natl. Acad. Sci. USA*, **113**, E7126–E7135.
50. van der Horst,S., Snel,B., Hanson,J. and Smeekens,S. (2019) Novel pipeline identifies new upstream ORFs and non-AUG initiating main ORFs with conserved amino acid sequences in the *5′* leader of mRNAs in *Arabidopsis thaliana*. *RNA*, **25**, 292–304.
51. Chang,K.S., Lee,S.H., Hwang,S.B. and Park,K.Y. (2000) Characterization and translational regulation of the arginine decarboxylase gene in carnation (*Dianthus caryophyllus* L.). *Plant J.*, **24**, 45–56.
52. Combier,J.P., de Billy,F., Gamas,P. *et al.* (2008) Trans-regulation of the expression of the transcription factor MtHAP2-1 by a uORF controls root nodule development. *Genes Dev.*, **22**, 1549–1559.
53. Leutz,A., Barbosa-Silva,A., Andrade-Navarro,M.A. and Wethmar,K. (2013) uORFdb—a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.*, **42**, D60–D67.
54. Kearse,M.G. and Wilusz,J.E. (2017) Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.*, **31**, 1717–1731.